LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Measuring and Exploiting Redundancy in Multi-Turn Reasoning Traces**

Supervisor: Ali Modarressi

Examiner: Hinrich Schütze

Level: MSc

Summary: Large Reasoning Models (LRMs) boost performance by generating a reasoning trace before the final answer, but in multi-round chat settings the model produces a new trace each turn while earlier traces are usually dropped (context limits, and typical single-trace training). This project asks whether consecutive traces contain substantial **semantic redundancy**—i.e., repeated constraints, intermediate conclusions, and stable subgoals—and whether preserving only that reusable content can improve later-turn performance.

We propose a two-part method. (1) **Quantify trace-to-trace semantic overlap** across turns by aligning reasoning traces from consecutive rounds using two complementary lenses: retrieval-style similarity (dense/sparse retrieval to match spans across traces) and model-based semantic matching (e.g., entailment/semantic equivalence scoring between extracted propositions). These tools are used to *measure and characterize redundancy* (what categories repeat, how overlap changes with dialogue length/task type), not as competing systems. (2) **Exploit redundancy** by extracting the stable/redundant information into a compact "reasoning memory" (e.g., constraints, derived facts, intermediate results) that is appended to the next-turn input, and test whether this improves correctness and consistency under fixed context budgets.

Evaluation will be conducted on established multi-round benchmarks—e.g., **MT-Bench** for general multi-turn assistant behavior and **CoQA** for conversational question answering or even a multi-step math setting to stress long dependency chains and repeated derivations across follow-up turns.

Requirements: vLLM (or SGLang), HuggingFace